# *Data Mining Techniques for Mortality at Advanced Age*

Lijia Guo, Ph.D., A.S.A. and Morgan C. Wang, Ph.D.
University of Central Florida

*Abstract*

This paper addresses issues and techniques for advanced age mortality study using data mining techniques, a new technology on the horizon with great actuarial potential. Data mining is an interactive information discovery process that includes data acquisition, data integration, data exploration, model building, and model validation. Both expert opinion and information discovery techniques are integrated together to guide each step in the information discovery process. Seven factors were considered in this study and the influences of these factors on advanced-age mortality distribution were identified with exploratory data analysis and decision tree algorithm. Models to address their effects on advanced age mortality were built with logistic regression technique. These models will be derived for projecting advanced age mortality distribution.

**Key Words:** *Mortality, Data Mining, Decision Tree, Logistic Regression, Data Exploration, and Information Discovery.*

## 1. Introduction

Modern technology and living has resulted in increasing numbers of people living into their 90s and beyond. Prospects of longer life have led to concern over their implications for public spending on old-age support and other related topics. It is necessary to better estimate advanced age mortality for assuring the solidity of government and private pension plans; for improving life insurance and annuity pricing; for designing and pricing long term care insurance, and other actuarial practice.

Mortality tables currently in use have used a variety of mathematical techniques to generate mortality rates for advanced ages as a smooth extension of the patterns of mortality rates of septuagenarians and octogenarians. Keyfitz (1968) reviews some earlier methods; Pollard and Streathfield (1979) summarizes more recent methods. Families of curves that provide a good fit to recent U.S. experience are employed by Heligman and Pollard (1980) and by McNown and Rogers (McNown 1992, McNown and Rogers 1989, 1992). The most popular model is the Gompertz law that assumes that the force of mortality continues to increase exponentially with age is used. Various modifications can be found at Pollard and Streathfield (1979) and Wilkins (1981). Tuljapurkar and Boe (1998) believe that Gompertz model "ignores the basic statistical

caution that a curve fitted over a particular domain is not necessarily a reliable guide outside that domain." There are several approaches used to develop a basic scientific theory of mortality including the evolutionary theory of senescence (Rose 1991, Tuljapurkar 1998), bio-actuarial theories (Pollard and Streathfield 1979), and hypotheses based on reliability theory (Gavrilov and Gavrilova 1991, Wachter and Finch 1997). All these approaches aim at explaining the age pattern of mortality. A comprehensive literature review is given by Tuljapurkar and Boe (1998).

Data mining is the process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules (Berry and Linoff, 2000). A typical data mining process includes data acquisition, data integration, data exploration, model building, and model validation. Both expert opinion and data mining techniques play an important role for each step of this information discovery process. However, this article focuses on data mining techniques and uses this study to illustrate a typical information discovery process.

In this study, we hypothesize that the mortality rates for advanced ages have different distribution among seniors with different backgrounds. Six factors that include gender, pay type (hourly and salaried), annuity size (large, medium, and small), participation status (beneficiary, disabled, retiree, employee, and combined), collar (blue, white, and mixed), and union or non-union member are considered. Instead of studying the mortality rates for whole advanced age seniors, we estimated the probability of mortality of each individual based on these factors under consideration since individual characteristics are responsible for mortality differences. We then combine these probabilities together to form the aggregate population mortality distribution for advanced ages.

The paper is organized as follows. Section 2 illustrates the information discovery process through data mining. Section 3 models the advanced age mortality by applying the data mining method with the SOA database for RP-2000 Mortality Tables. This section identifies factors that influence mortality at advanced ages and quantify their effect and interactions. It also discusses problems with data quality found in the data database and techniques for dealing with these problems. Some issues and techniques for projecting advanced age mortality improvement are also presented in this section. Section 4 concludes, offering some prospective thought for the future study.

## 2. Data Mining

Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in models and patterns that are both understandable and useful to the data owner (Hand, Mannila, and Symth, 2001). It is the process of non-trivial extraction of implicit, previously unknown and potentially useful information such as knowledge rules, constraints, and regularities from data stored in repositories using pattern recognition technologies as well as statistical and mathematical techniques.

A typical data mining process includes data acquisition, data integration, data exploration, model building, and model validation. This article focuses on using data mining techniques to illustrate the information discovery process for the old age mortality model.

The input to a data mining algorithm is a *training set* of records, each of which is tagged with a class label. A set of attribute values defines each record. Attributes with discrete domains are referred to as *categorica*l, while those with ordered domains are referred to as *numeri*c. The goal is to induce a model or description for each class in terms of the attributes. The model is then used to predict future records whose classes are unknown.

The popular data mining techniques include Bayesian analysis, (Cheeseman et al., 1988), neural networks (Bishop, 1995; Ripley, 1996), genetic algorithms (Goldberg, 1989), decision trees (Breiman et al., 1984), and logistic regression (Hosmer and Lemeshow, 1989). We propose to use a hybrid method that combines the strength of both logistic regression and decision trees in this paper.

Logistic regression is a special case of the generalized linear model. It has been used for studying the odds ratio for a very long time and its properties have been well studied by statistical community. Easy to interpret is one advantage of modeling with logistic regression. Assume that the data set consist of i = 1, 2, …, n records. Let $p_i, i = 1, 2, \hbar\ , n$ be the corresponding mortality rate for each record and $x_i = (x_{1i}, x_{2i}, \hbar\ , x_{ki})$ be a set of k variables associated with each record. A linear-additive logistic regression model can be expressed as

$$\text{logit} = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji}, \textit{ where } i = 1, 2, \hbar\ , n.$$

If the model is correctly specified, each dependent variable affects the logit linearly. Exponentiation the parameter estimate of each slope gives the odds ratio, which compares the odds of the event of one category to the odds of the event in another category. However, it poses several drawbacks especially when the size of the data getting large. The curse of dimensionality makes the detection of nonlinearities and interactions difficult. If the model is not correctly specified, the interpretation of the model parameter estimates becomes meaningless. In addition, the data might not be evenly distributed among the whole data space. It is very likely that some segments of the data space have more records than other segments. One model fits the whole data space might not be the best choice. Although there are many existing methods such as backward elimination and forward selection that can help data analyst to build logistic regression model, blindly apply these methods might not be the best choice. The proposed hybrid method uses the decision trees to identify non-linearity and interactions in high dimensions for each segment of the data space that can guide the data analyst to obtain the best model in each data segment.

The decision tree is efficient and is thus suitable for large data sets. Decision trees are perhaps the most successful exploratory method for uncovering deviant data structure.

Trees recursively partition the input data space in order to identify segments where the records are homogeneous.

Decision trees are tree-shaped structures that represent sets of decisions that either:

- Classify or predict observations.
- Predict the appropriate decision when you specify decision alternatives.

When you classify or predict observations, you classify values of nominal or binary targets. For interval targets, you can predict outcomes. Trees produce a set of rules that can be used to generate predictions for a new data set. These rules can also be used to detect interactions among variables and to decide the importance of a specific variable.

Specific decision tree methods include Classification and Regression Trees (CART; Breiman et. al., 1984) and the count or Chi-squared Automatic Interaction Detection (CHAID; Kass, 1980) algorithm. CART and CHAID are decision tree techniques used for classification of a data set. In this article, we use CHAID like tree to guide us on variable selection and interaction detection because all variables investigated are classification variables except the variable age. Although decision tree can split the data into several homogeneous segments and the rules produced by tree can be used to detect interaction among variables, it is relatively unstable and it is difficult to detect linear or quadratic relationship between the response variable and the dependent variables. For example, the tree will produce a step function to explain the relationship among age and mortality rate even if the true relationship between age and mortality rate is linear or quadratic. Thus, some preliminary work needs to be done before using the decision tree to guide us to discover the data segments and correct model for each segment.

Therefore the proposed hybrid method includes three steps. The first step is to identify the importance of the risk factors in determining the advanced age mortality distribution.. We fit a logistic regression model using only age and square term of age as the depend variables because it is well know that mortality rate are grow exponentially with age. If the mortality rate only depends on age and square term of age, the residuals of this logistic regression model should not have any special pattern. If there are other variables that need to include in the model, the decision trees should be able to identify them through these residuals. Thus, the next step of this hybrid method is using decision trees on these residuals. In this study, the decision tree algorithm identifies six segments and picks up different set of variables and interaction terms in each segment. These segments will be discussed in details in Section 3.2. Finally, a logistic regression model for each segment is developed. All the important variables and important interaction terms are included in each segment logistic regression model. Since the number of records available in each segment is different, important variables and interaction terms selected in each segment are different in this study. Consequently, the predicted power for each segment logistic regression model is different. Detail description of each segment's logistic regression model is given in Section 3.2 as well.

## 3. Modeling of Mortality in Advanced Age

This section models the advanced age mortality using the decision tree techniques. The database will be used for this study is the SOA database for RP-2000 Mortality Tables. Section 3.1 reviews the basic theory in advance age mortality study and discuss the data. Section 3.2 derives the model and discusses model properties in the context of data mining. Section 3.3 analyzes the outcomes of the model.

### *3.1 Theories of Mortality and the data*

The theoretical basis for mortality analysis includes the traditional actuarial methods for life table analysis, the curve-fitting methods based on experience studies, and more fundamental arguments (for example, genetic, evolutionary, reliability theory) about the nature of mortality, and dynamic theories of mortality in populations.

The current mortality tables analysis uses a variety of mathematical techniques to generate mortality rates for advanced ages as a smooth extension of the patterns of mortality rates of septuagenarians and octogenarians. The traditional actuarial method uses the graduation of mortality curves, to represent the age pattern of mortality $\mu_x$ by an analytical law. Keyfitz (1968) reviews some earlier methods; Pollard and Streathfield (1979) summarizes more recent methods. Heligman and Pollard (1980), and McNown and Rogers (McNown 1992, McNown and Rogers 1989, 1992) study the families of curves that provide a good fit to recent U.S. experience. Among all the analytic mortality laws, Gompertz law ( $\mu_x = Bc^x, B > 0, c > 1$ for all age *x*) has been the most popular curve used to model the mortality distribution of all ages. For advanced age mortality study, however, as pointed out by Wilkins (1981), Gompertz law has its limitation. Tuljapurkar and Boe (1998) believe that Gompertz model "ignores the basic statistical caution that a curve fitted over a particular domain is not necessarily a reliable guide outside that domain." Some researchers believe that mortality rates level off to a plateau where the mortality rate is constant but less than certain for very high ages. Still others believe that mortality rates decrease after reaching a maximum rate.

Wachter and Finch (1997) gave a comprehensive review of some advanced study on the mortality theory. These approaches aim at explaining the age pattern of mortality. For example, the population genetic methods that underlie the evolutionary approach hold some promise of integrating new information about the genetic components of disease and mortality (Scriver 1984, Wachter and Finch 1997, Weiss 1993). Another approach is based on the reliability theory that generates the failure rate of multi-component systems and focus on the prediction of age patterns of mortality (see Gavrilov and Gavrilova, 1991). Tuljapurkar and Boe (1998) review the all these approaches and the comparison of the methods.

Similar to the dynamic models of mortality developed in last two decades (see Vaupel et al. 1979, Vaupel and Yashin 1985, Vaupel et al. 1988, Yashin et al. 1985, Manton 1991, Manton and Stallard 1994, Manton et al. 1993, 1994, 1988), the approach proposed in this paper, analyzes the individual characteristics that determine mortality differences (the risk factors) and how they affect the aggregate population mortality for advanced age.

The database used for this study is the SOA database for RP-2000 Mortality Tables. The data set included 10,957,103 exposed life-years and 190,928 deaths. We used the subset of the database that includes all the lives above age 70. The risk factors listed by the data include age, gender, participation status, union, pay type, collar type, and annuity amount, etc. Upon data availability, more variables could be analyzed using our model and more accurate estimates for advanced age mortality should be yielded.

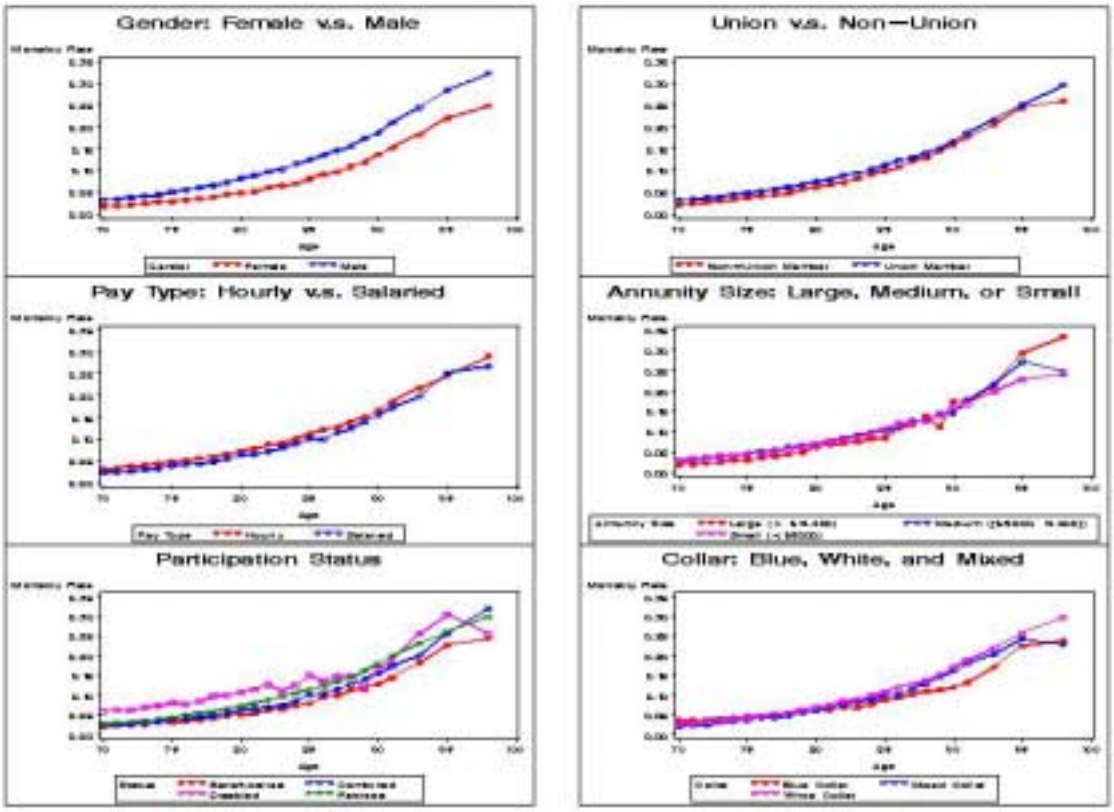
### *3.2 Data Mining Modeling*

We now start the modeling process by studying the relationship between mortality and the underlying risk factors including age, gender, pay type, union, annuity size, participation status, and collar. A hybrid method is developed for this study – the modeling process is a combination of the decision tree techniques and logistic regression.

First, we use decision tree algorithm to identify the factors that influence mortality. After the factor being identified, the observations are grouped into several segments and the logistic regression technique is used to quantify the mortality effect for each segment. Instead of estimating the mortality rate for all advanced age seniors, there is an estimated mortality for each group identified by the tree algorithm.

Our exploratory data analysis shows the effects of demographic factors such as age, gender, participation status, union, pay type, collar type, and annuity amount on the mortality at advanced age illustrated in Figure 1.

## Figure 1 Advanced Age Mortality Trend




As indicated in Figure 1, the main effect exists for all six variables considered. Our analysis reveals that the mortality rates changes rather smoothly between age 70 and 85 and increasing sharply (exponentially growth) beyond age 85 for both male and female. The analysis also shows, however, that the degrees of the effects of the risk factors including participation status, union, pay type, collar type, and annuity amount on male and female are different. A reliable mortality model should consider all the risk factors, the interaction of these factors as well as the importance of the factors.

We now use the decision tree algorithm to analyze the influences and the importance of the mortality risk factors. The tree algorithm used in this research is SAS/Enterprise Miner Version 4.2 (2001). We built 100 binary regression tree s and 100 CHAID like trees and our decision tree analysis reveals that the participant status has the greatest impact on the mortality rate (see Table 1). The mortality, and the interaction among different factors which affect the mortality, vary as the participant status changes. Further more, there is a significant gender influence within the "Combined" status.

Based on this analysis, we build our mortality model for each of the following six segments: Employees, Beneficiaries, Combined, Disabled, Male Retirees, and Female Retirees. We developed the mortality model using logistic regression method for each of the segments. The results are detailed in the next section.

Table 1 Variable Importance Measure

| Variable | Importance |
|---|---|
| Participation Status | 1.00 |
| Gender | 0.75 |
| Annuity size | 0.43 |
| Pay Type | 0.21 |
| Union | 0.18 |
| Collar | 0.00 |

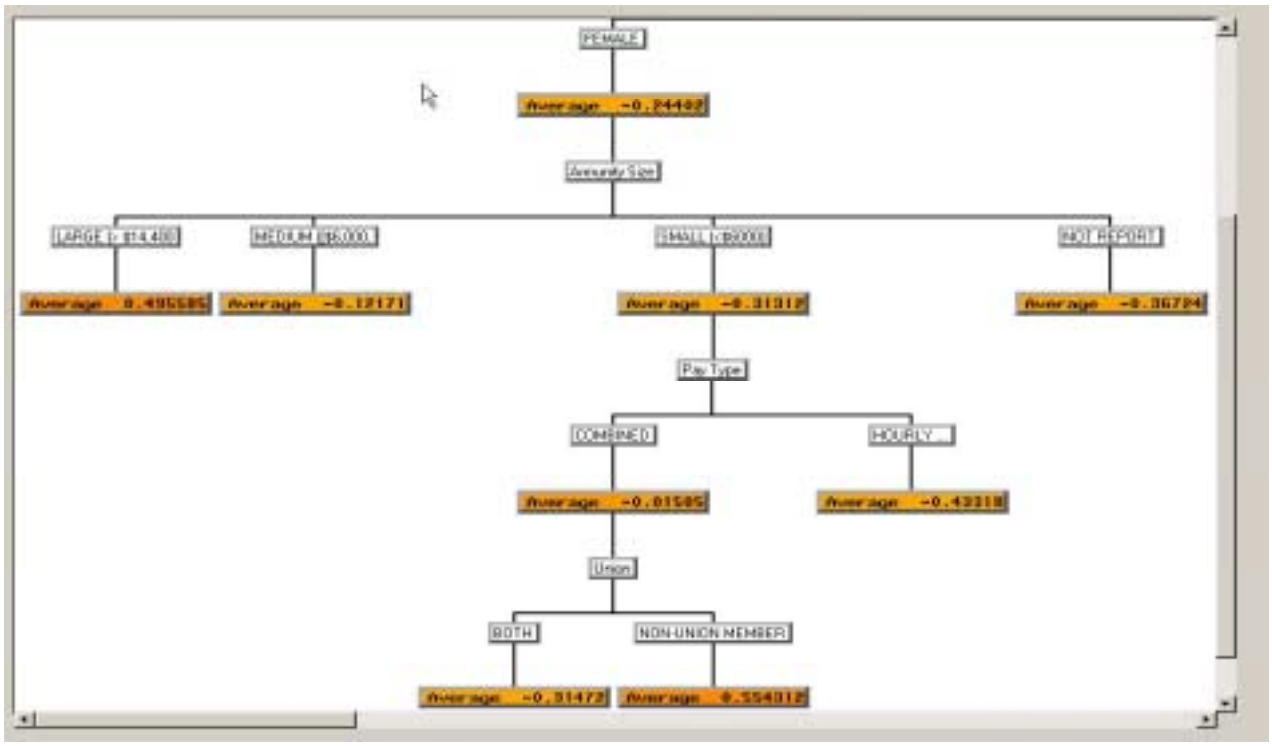
### *3.3 The results*

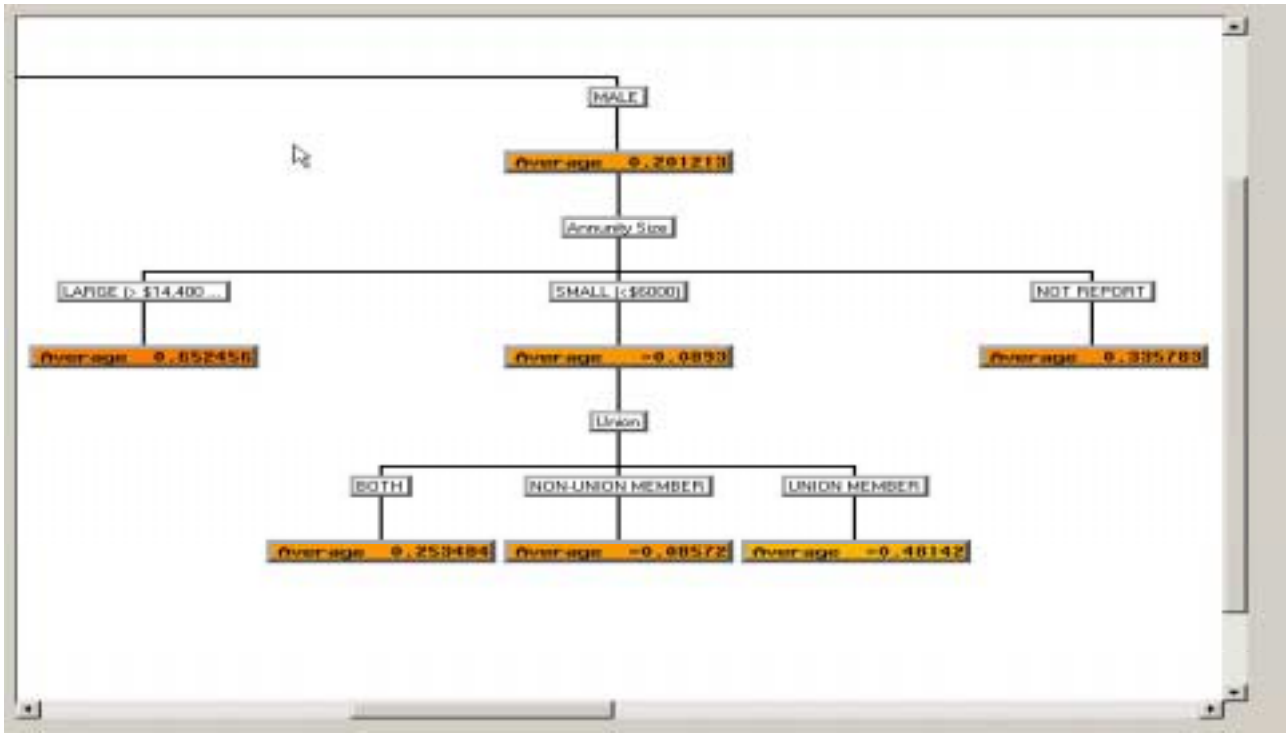
We now discuss the mortality models for each of the six segments.

I. “Beneficiaries” Segment

Let begin with the segment for the “Beneficiaries” status. A tree analysis is applied for the other mortality factors. From the Figure 2 below, one can see that the most important variable in this segment is Gender and the next important variable is “Annuity Size”. Since the within node mean squares errors do not have similar pattern among female and male, the interaction among “Gender” and “Annuity Size” does exist. In addition, the variable “Union” is important for female and the variable “Pay Type” is important for male.

# Figure 2 Tree Diagram for Beneficiaries



The tree analysis clearly implies that the mortality rate for this segment should be related to the following variables and their interactions: age, gender, annuity size, union, pay type.

Notice that, the variable "Collar" is not selected by any tree in above tree diagram. This implies that there is no significant difference among log odds ratio between different categories in "Collar". Therefore, the model ignores the variable "Collar".

Applying the standard Logistic Regression analysis, the final mortality model is

$$\log\left(\frac{p}{1-p}\right) = 0.10 - 0.15x + 0.0015x^2 - 0.48I_G + S + T + U + ? \qquad (1)$$

Where

$p$ is the mortality rate;

$x$ is the age;

$I_G$ is the indicate variable for gender: $I_G$=1 for female and $I_G$ =0 for male;

$S$ is the factor that represents the annuity amount:

$$S = \begin{cases} 0.87 & large\ annuity \\ -0.049 & median\ annuity \\ -0.55 & small\ annuity \end{cases};$$

$T$ is the pay type factor, where

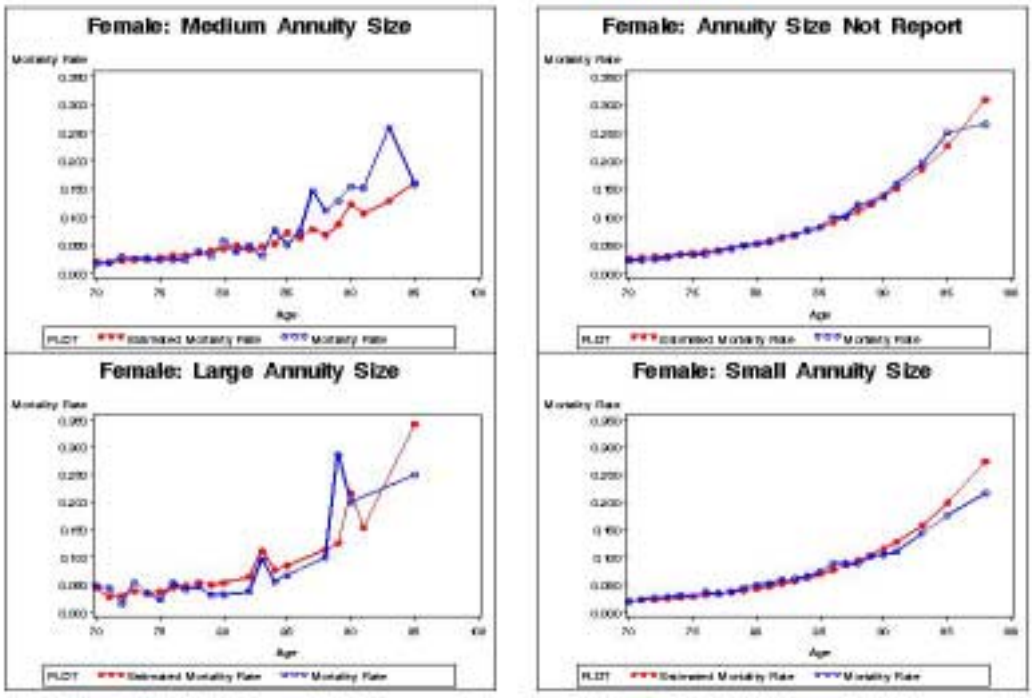
$$T = \begin{cases} -0.18 & hourly\ pay \\ 0 & salaried\ pay \\ 1.02 & combined \end{cases};$$

$U$ is the pay type factor, where

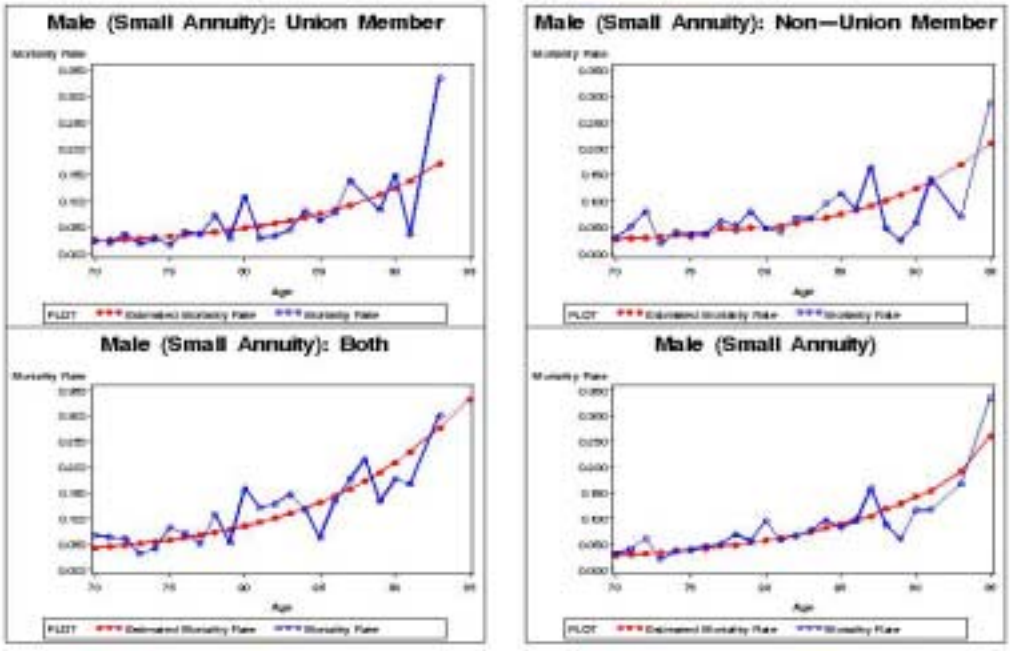
$$U = \begin{cases} 0.52 & nonunion \\ 0 & union \\ -0.82 & combined \end{cases}.$$

The comparison of the estimated mortality to the historical mortality rates is displayed in Figure 3 and Figure 4.

## Figure 3 Mortality Model for Female (Beneficiaries)




## Figure 4 Mortality Model for Male (Beneficiaries)

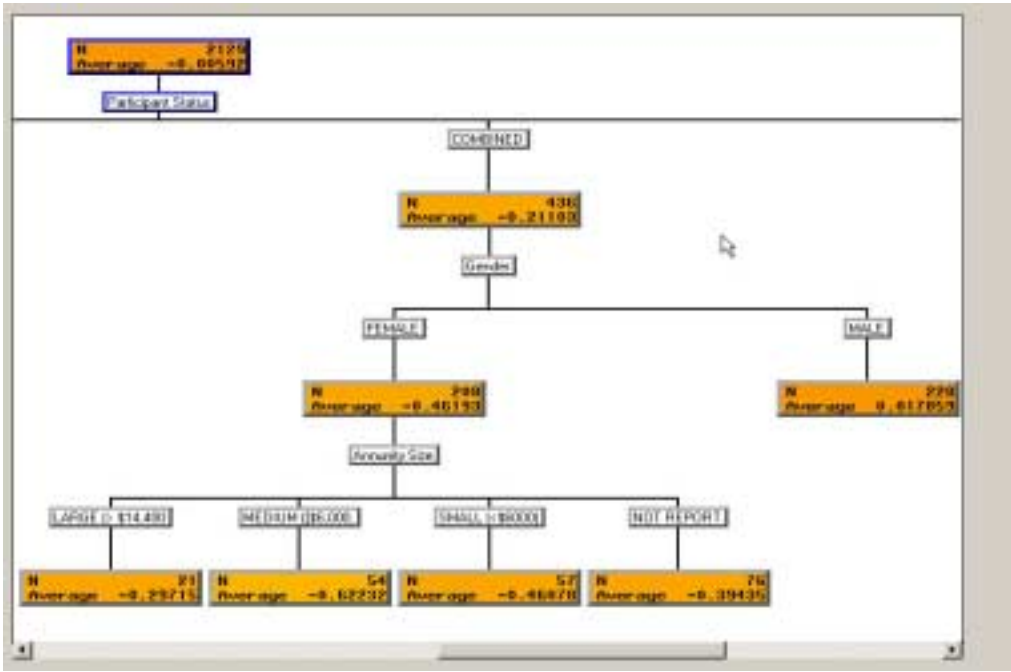



The R-square of the regression model is 0.72.

II. “Combined” Segment

Next, we look at the segment for “Combined” status. From Figure 5, one can see the most important variable is “Gender” and the next important variable is “Annuity Size”. Although “Annuity Size” seems not important for male, the model for this segment should include both “Annuity Size” and Gender. Since the ‘annuity Size’ is not important for male, this is an indication for the interaction between ‘Gender’ and ‘Annuity Size’. Both “Pay Type” and “Union” do not appear on any part of the tree. This implies that they are not significant when study the log odds ration between categories. Finally, there is few data in the “Collar” category.

Figure 5. Tree Diagram for the “Combined” Segment




Therefore, the mortality distribution for this segment is determined by “age”, “ gender”, “annuity size”, and their interactions.
Applying the Logistic regression analysis, the mortality model is derived as:

$$\log\left(\frac{p}{1-p}\right) = -7.66 + 0.013x + 0.00058x^2 - 0.22I_G + S + GS * I_G \ , \qquad (2)$$

Where

$p$ is the mortality rate;
$x$ is the age;
$I_G$ is the indicate variable for gender: $I_G$=1 for female and $I_G$ =0 for male;
$S$ is the factor that represents the annuity amount:

$$S = \begin{cases} -0.062 & large\ annuity \\ -0.066 & median\ annuity \\ 0.033 & small\ annuity \end{cases};$$

$GS$ is the factor that indicates the interactions:

$$GS = \begin{cases} 0.13 & large\ annuity \\ -0.0091 & median\ annuity \\ -0.087 & small\ annuity \end{cases}.$$

The forecast of the mortality trend for the "Combined" participating status is given in the following figures.

**Figure 6 Mortality Model for Male (Combined)**

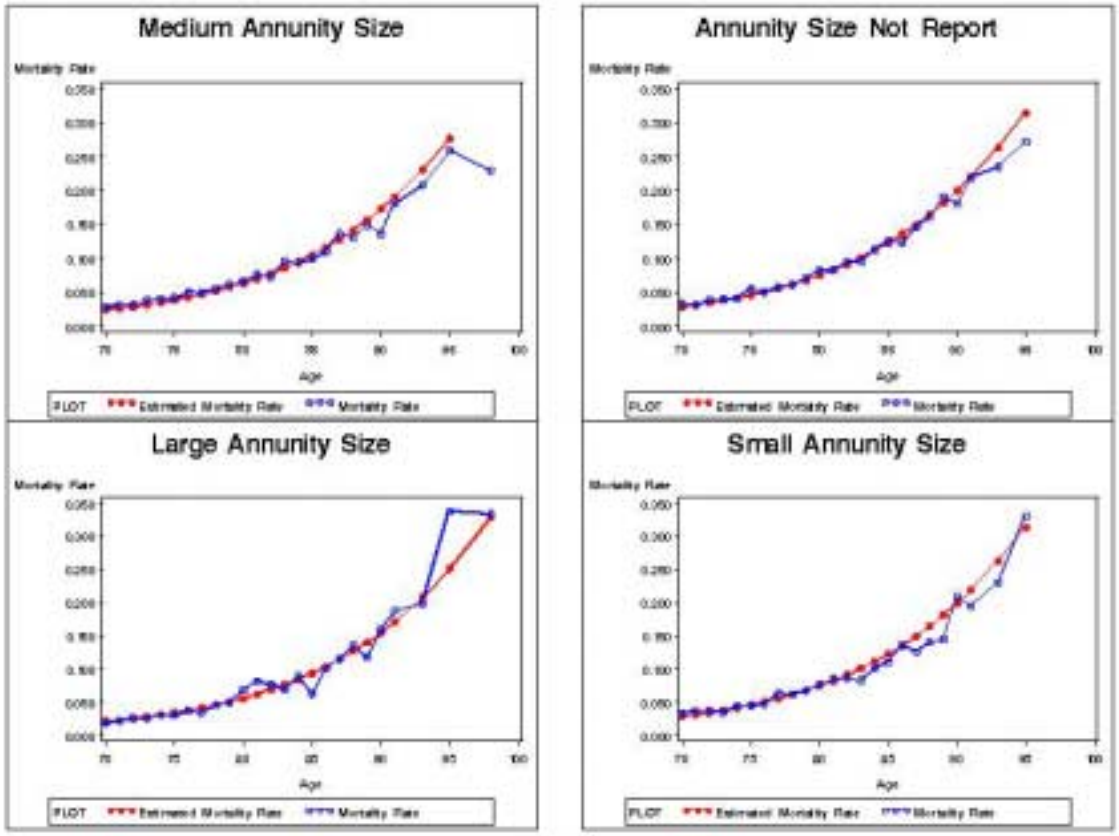
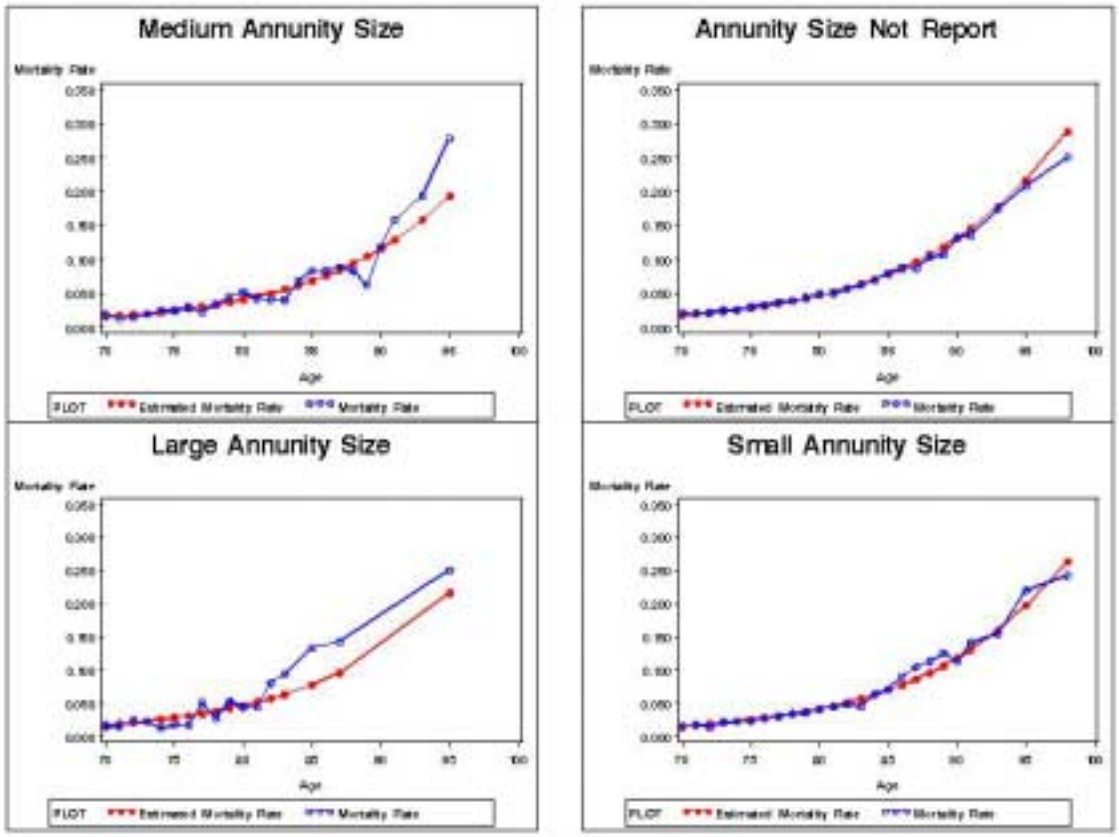


**Figure 7 Mortality Model for Female (Combined)**




The Figure shows a nice fit of the model and the fitness of the model is given in next table:

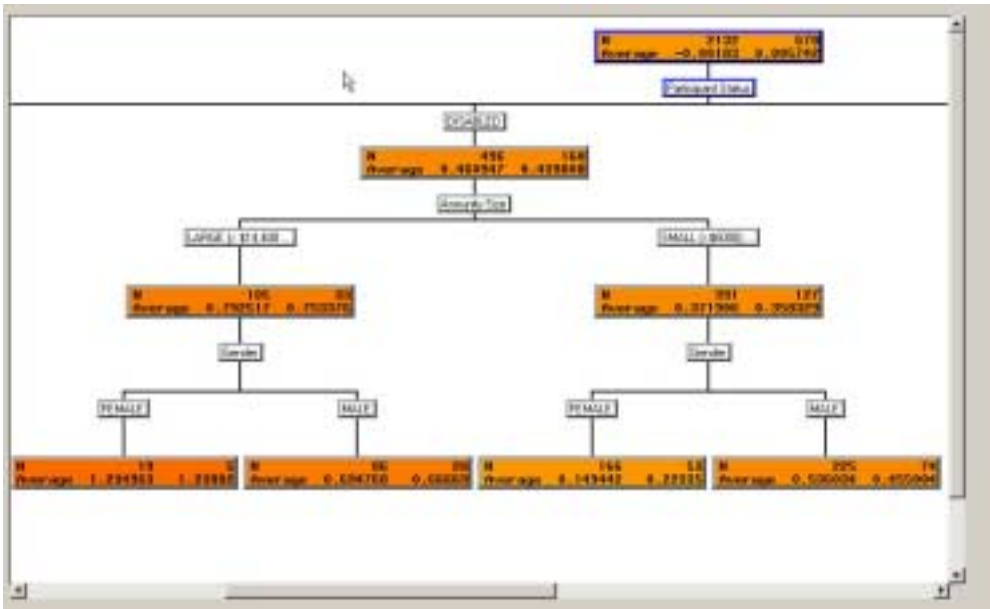
Model Fitting Information

| | | | |
|---|---|---|---|
| R-square | 0.96 | Adj R-sq | 0.95 |
| AIC | -613.15 | BIC | -610.05 |
| SBC | -580.62 | C(p) | 10.00 |

III. "Disabled" Segment

The tree analysis on the data structure for the "Disabled" segment indicates that the most important variable is "Annuity Size" and the second important variable is "Gender" as shown in Figure 8. There are interaction among 'Annuity size' and 'Gender' because the average mean square error for female is larger for annuity size 'Large' and 'Medium' and it is smaller for annuity size 'Small' and 'Not Report'.

Figure 8. Tree Diagram for the "Disabled" Segment



The Logistic Regression model, with R-square value 0.6714, is

$$\log\left(\frac{p}{1-p}\right) = -4.73 - 0.015x + 0.005x^2 - 0.053I_G + S + GS * I_G \qquad (3)$$

where

$p$ is the mortality rate;

$x$ is the age;

$I_G$ is the indicate variable for gender: $I_G$=1 for female and $I_G$ =0 for male;

$S$ is the factor that represents the annuity amount:

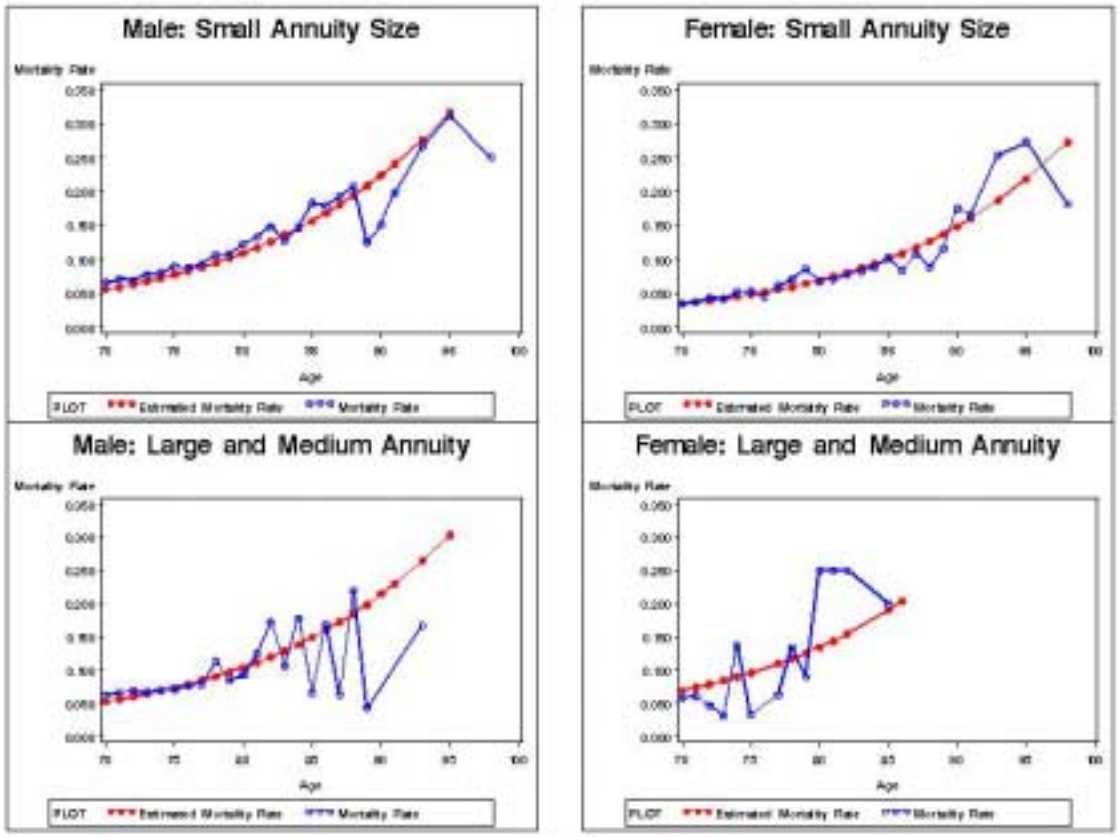
$$S = \begin{cases} 0.17 & large\ annuity \\ 0.17 & median\ annuity \\ 0 & small\ annuity \end{cases};$$

$GS$ is the factor that indicates the interactions:

$$GS = \begin{cases} 0.2 & large\ annuity \\ 0.2 & median\ annuity \\ 0 & small\ annuity \end{cases}.$$

And the effect of the model is given by

## Figure 9 Mortality Model for Disabled




We note that the effectiveness of the model is reduced for both male and female with large and medium annuity due to the limited information in the data.

IV. “Employee” Segment

For the “Employee” segment, we built 100 binary regression tree s and 100 CHAID like trees and analyzed contributions of all five variables to the mortality rate.

Since the only Category remained for annuity size is ‘Not Report’, so the annuity size is not considered in the model. The tree does not select “Union” nor “Collar”. The odds ratios among different categories for both “Union” and “Collar” are not significantly different. Also, there are very few data in categories such as ‘Union’ and ‘Both’; ‘Blue collar’ or ‘White Collar’. Therefore, both “Union” and “Collar” variables are not considered in the mortality model.

We notice an interesting fact that the “Gender” variable is not selected by tree and there are very few data of female over 77. Table 2 shows the Odds ratio of Female to male mortality.

Table 2. Odds Ratio of Female to Male Mortality.

| Age | Odds | Lower | Upper |
|---|---|---|---|
| 70 | 0.777 | 0.371 | 1.627 |
| 71 | 1.074 | 0.381 | 3.028 |
| 72 | 0.648 | 0.195 | 2.159 |
| 73 | 0.240 | 0.013 | 4.482 |
| 74 | 0.896 | 0.265 | 3.031 |
| 75 | 0.425 | 0.073 | 2.480 |
| 76 | 1.021 | 0.149 | 7.017 |
| 77 | 2.679 | 0.274 | 26.219 |
| 78 | 0.568 | 0.027 | 12.061 |
| 79 | 3.545 | 0.069 | 182.80 |
| 80 | 1.713 | 0.564 | 5.206 |
| 81 | 1.456 | 0.055 | 38.594 |
| 82 | 7.545 | 0.136 | 420.10 |
| 83 | 3.762 | 0.128 | 110.72 |
| 84 | 8.143 | 0.138 | 478.77 |
| 85 | 2.143 | 0.072 | 63.758 |
| 86 | 8.600 | 0.138 | 536.30 |
| 87 | 9.667 | 0.136 | 688.05 |
| 90 | 7.667 | 0.107 | 550.13 |
| 91-92 | 2.730 | 0.052 | 142.62 |
| 93-94 | 2.556 | 0.048 | 135.42 |

Further analysis shows that the 95% confidence intervals for odds ratio of each age group contain 1. This means the mortality rate for female is not significant lower than the rate for male.

Finally, we examine the “Pay type” variable. Again, “Pay type” is not selected by the tree algorithm. Table 3 shows that although the odds ratios are consistently lower than 1 for age over 80, there is very few data for age over 80.

Table 3 Odds Ratio (Hourly v.s. Combined)

| Age | Odds | Lower | Upper |
|---|---|---|---|
| 70 | 1.135 | 0.555 | 2.322 |
| 71 | 2.509 | 0.597 | 10.539 |
| 72 | 0.964 | 0.251 | 3.711 |
| 73 | 1.035 | 0.136 | 7.886 |
| 74 | 0.372 | 0.100 | 1.374 |
| 75 | 2.753 | 0.668 | 11.350 |
| 76 | 1.237 | 0.160 | 9.543 |
| 77 | 0.405 | 0.019 | 8.580 |
| 78 | 1.409 | 0.027 | 72.277 |
| 79 | 1.319 | 0.026 | 68.141 |
| 80 | 0.301 | 0.017 | 5.268 |
| 81 | 0.257 | 0.004 | 14.826 |
| 82 | 0.297 | 0.005 | 16.791 |
| 83 | 0.355 | 0.006 | 20.141 |
| 84 | 0.333 | 0.006 | 20.161 |
| 85 | 0.333 | 0.005 | 21.638 |
| 86 | 0.200 | 0.003 | 14.653 |
| 91-92 | 2.600 | 0.047 | 143.29 |

The 95% confidence intervals for odds ratio of each age group contain 1 when compare 'Hourly' category against 'Combined' category. This means the mortality rate for both categories is not significantly different among the pay types.

Based on the above analysis, the mortality model for "Employee" status should only depend on the "age" factor.

The logistic Regression Model is

$$\log\left(\frac{p}{1-p}\right) = 21.79 - 0.77x + 0.01x^2 \qquad (4)$$

and the R-square for this fitting is 0.4917.

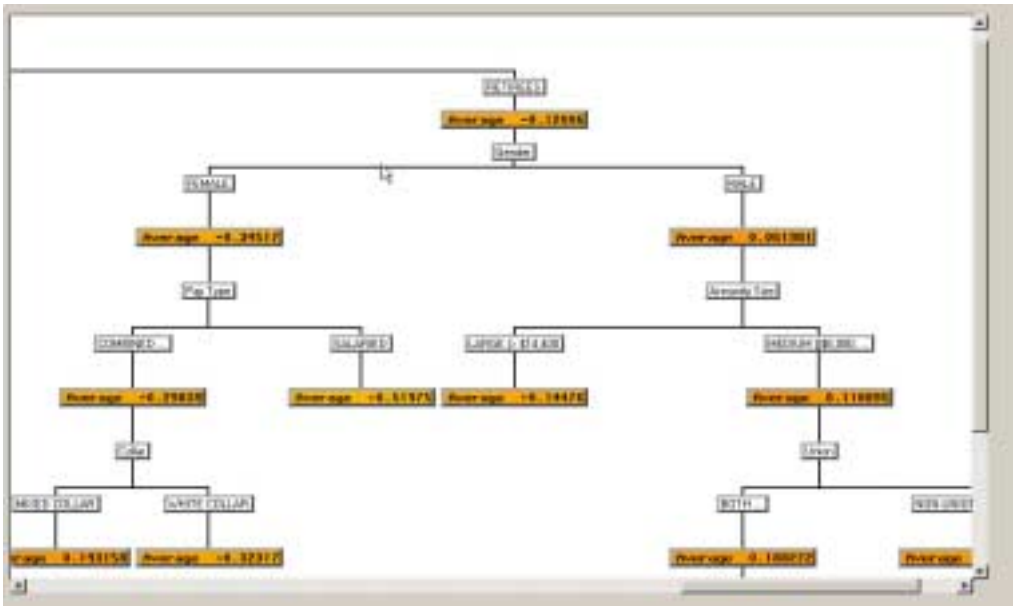
We note that because of the lack of the data in this segment especially for age above 85, the model is not very reliable for age older than 85.

V. Female "Retiree" Segment

Tree Analysis for the "Retiree" Segment is shown in Figure 10. From Figure 10 below, one can see that "Collar" and "Pay Type" are two important variables in determine the female mortality rates. Since the variable only selected by "Pay Type" = 'Combined", the interaction between "Collar" and "Pay Type" does exist. Both "annuity size" and "union" are not picked up by tree algorithm. Therefore, the female mortality distribution

should be a function of age, collar, pay type, and the interaction between collar and pay type only.

Figure 10. Tree Diagram for the "Retiree" Segment



Thus, the logistic regression model form for this segment is

$$\log\left(\frac{p}{1-p}\right) = -17.97 + 0.26x - 0.00087x^2 - C + PT - 0.046PTC, \qquad (5)$$

Where

$p$ is the mortality rate;

$x$ is the age;

C is the collar variable:

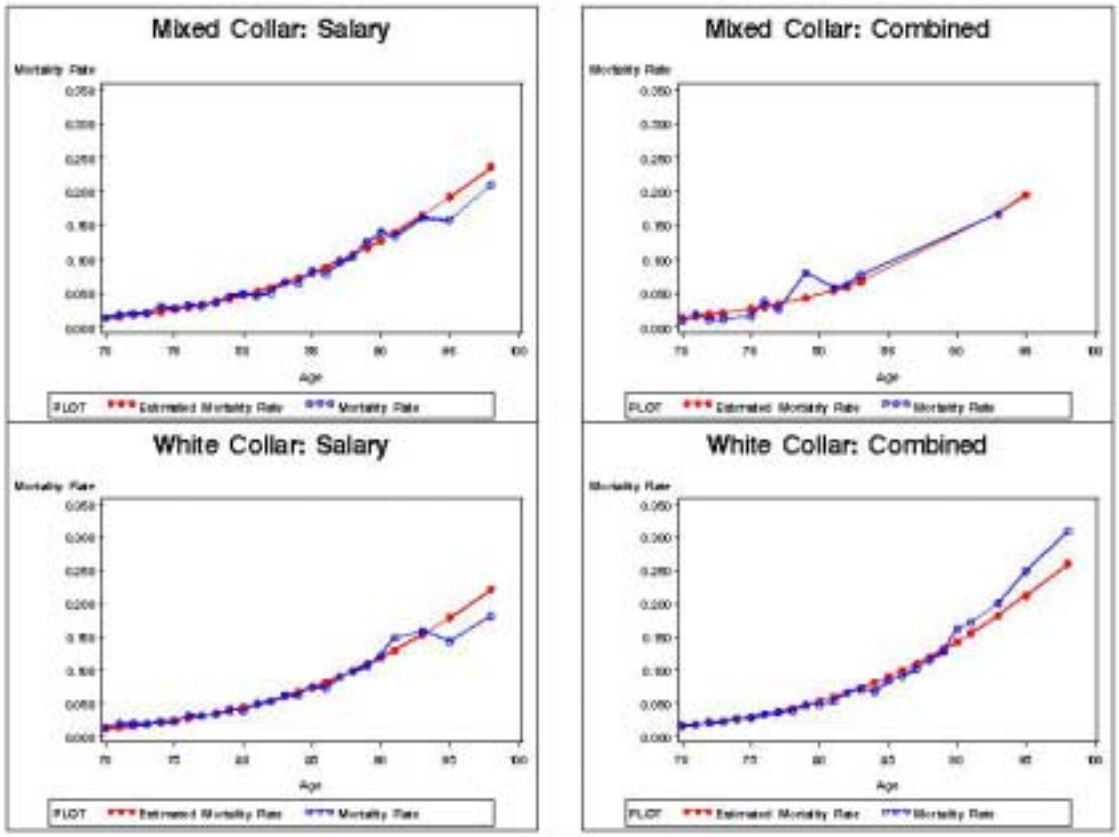
$$C = \begin{cases} 0 & white\ collar \\ 0 & blue\ collar \\ 0.0047 & mixed\ collar \end{cases}$$

$PT$ is the variable that represents the pay type:

$$PT = \begin{cases} 0.033 & combined\ pay\ type \\ 0.051 & hourly\ pay\ type \\ 0 & salarized\ pay\ type \end{cases}.$$

The R-square for the regression is 0.95 and the mortality model is displayed in Figure 11.

## Figure 11 Mortality Model for Female Retiree




VI. Male "Retiree" Segment

Unlike the female retirees, Figure 10 indicates that "Annuity Size" and "Union" are the important factors for modeling male retirees mortality. Since the variable only selected by "Annuity Size" = 'Medium", the interaction among "Annuity Size" and "Union" does exist. Thus, the mortality should be modeled by the age, annuity size, union, and the interaction between annuity size and union.

$$\log\left(\frac{p}{1-p}\right) = -14.57 + 0.20x - 0.00055x^2 - S - U + SU\,, \qquad (6)$$

Where

$p$ is the mortality rate;

$x$ is the age;

$S$ is the factor that represents the annuity amount:

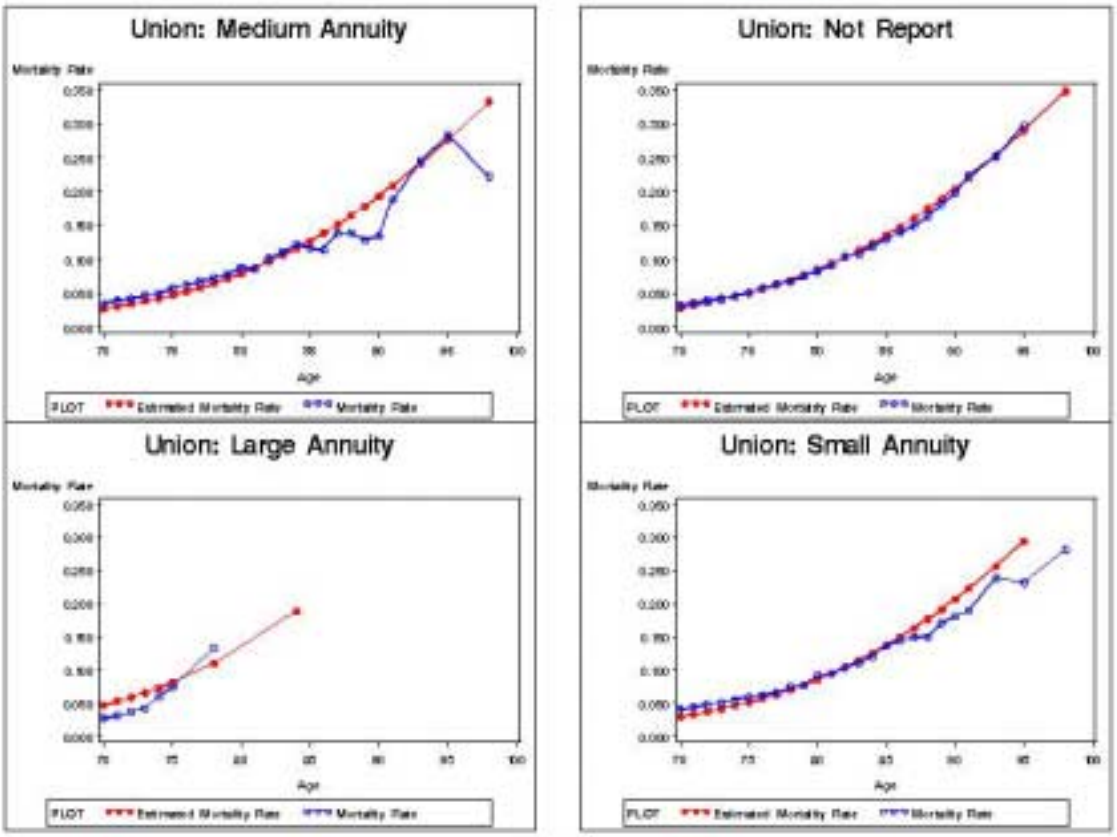
$$S = \begin{cases} 0.044 & large\ annuity \\ 0.060 & median\ annuity \\ 0.0074 & small\ annuity \end{cases};$$

$U$ is the variable that represents the union status:

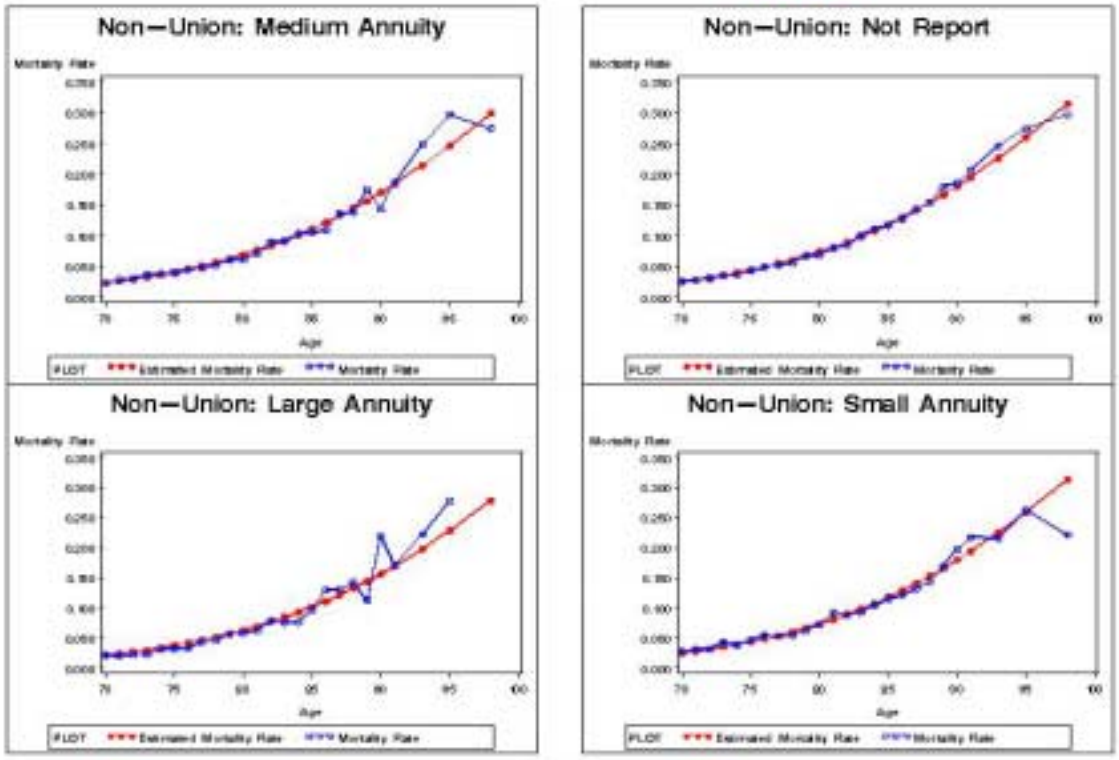
$$U = \begin{cases} 0 & union\ member \\ 0.14 & non-union\ member \\ 0.040 & combined \end{cases};$$

The R-square for this regression model is 0.92. Figure 12, Figure 13, and Figure 14 show the mortality models for male retirees and how the mortality estimates are affected by the annuity size variable, the union variable, and their interactions.

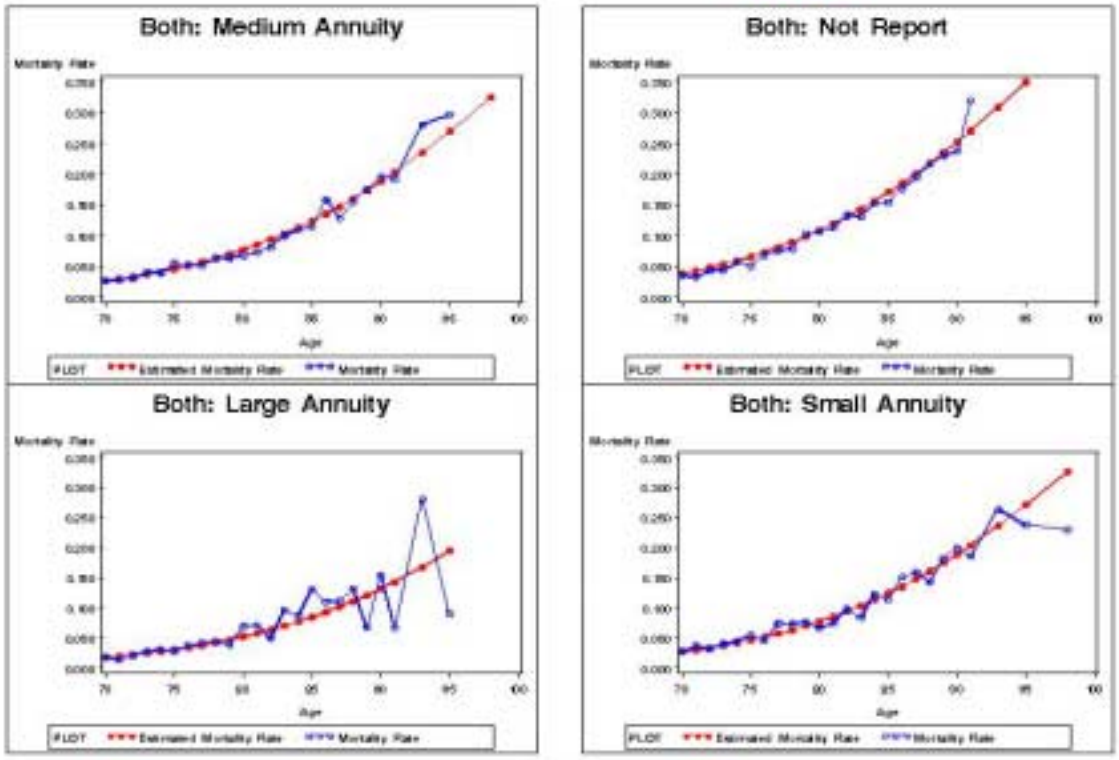
**Figure 12 Mortality Model for Male Retiree**



## Figure 13 Mortality Model for Male Retiree




## Figure 14 Mortality Model for Male Retiree



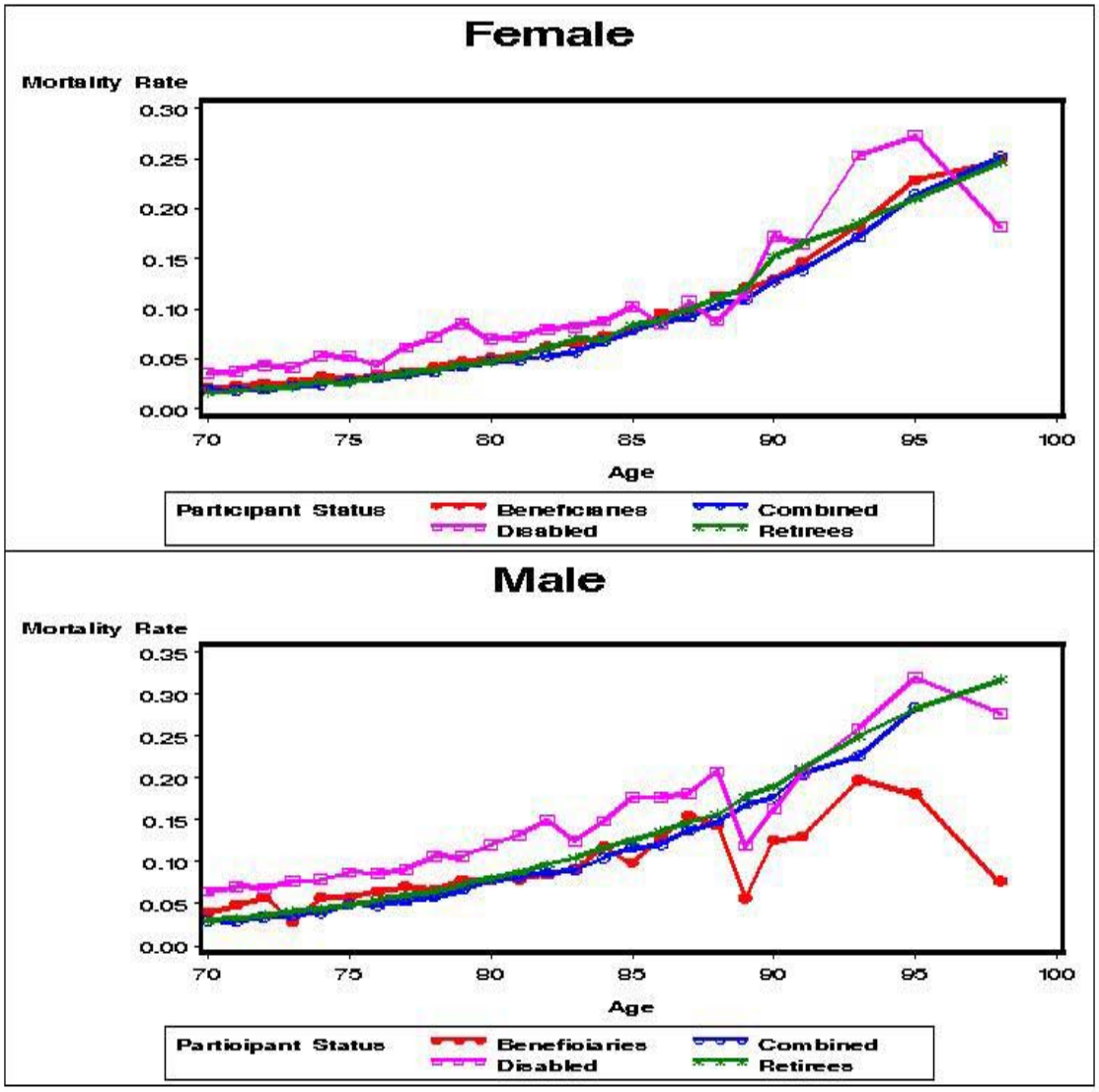
## 4. Conclusions

In this paper, we summarize the preliminary results of our study on the advanced age mortality using a hybrid data mining technique. We examined several risk factors for old age mortality. The influences and the correlations of these factors on advanced-age mortality distribution were identified with exploratory data analysis and decision tree algorithm.

Several interesting discoveries have been made while comparing the decision trees analysis results, some of them expected and others surprising and worth thinking over and doing further research.

The most significant observation described in this preliminary work is definitely the evidence of the "non Gompertz" mortality distribution between age 70 and 85. Figure 15 below describes the mortality rates by Participant Status for male and female separately.

# Figure 15 Advanced Age Mortality Trend




We notice that, from above figures, the Figure 4-1 in Comparison of RP-2000 Mortality Rates by Participant Status Females Ages 50 – 69, SOA THE RP-2000 MORTALITY TABLES; and the Figure 4-2 in Comparison of RP-2000 Mortality Rates by Participant Status Females Ages 50 – 69, SOA THE RP-2000 MORTALITY TABLES (http://www.soa.org/research/rp2000.html), the mortality rate show exponentially growth from age 50 to 69, linear growth between age 70 and 85, and then exponentially growth again for ages 86 and above.

The major difference between existing advanced age mortality techniques and hybrid data mining method presented can be summarized as follows. Many factors affect the mortality at advanced age. When the demographic of our seniors changed, the mortality

distribution of these seniors changed as well. Note the interaction of the variables and the importance analysis reveals that the selection of the risk factors may influence the quality of the mortality model.

This study uses data mining techniques to illustrate the information discovery process for the old age mortality model. As shown in Section 3.3, the mortality models derived by our hybrid data mining method varies with the most important risk factor (the participating status, in this study) among all the other variables. For example, the mortality of the beneficiaries is determined by gender, annuity size, the pay type, and their interactions (see equation (1) in Section 3.3). Another interesting discovery is that the male retirees mortality model and the female retirees mortality model depend on different variables. (see equation (5) and (6)). Based on the analysis, the collar and pay type are two irrelevant factors for male retiree mortality rate and should not be included in the male mortality model: Male retiree mortality depends on age, annuity size, union, and their interactions (see equation (6)). Meanwhile, for the female retiree mortality model, both the annuity size and union status are insignificant risk factors based on the current database. We believe as the female demography changed in the past three decade, variables such as annuity size, and union will play more important role in determining the female mortality. The gender factors will play a much-reduced role in determining beneficiaries' mortality model.

Due to the limitation of the database, our analysis show limited results on the mortality distribution for the ages above 95. Discuss problems with data quality found in the database and techniques for dealing with these problems will be one of the fields for future study. Issues and techniques for projecting advanced age mortality improvement will be available for further research.

There are other risk factors that determines an individuals mortality such as education, life style (smoking/non-smoking, for example) that are not provided in the database we used for our model. Upon the availability of the database, great improvement from our further research and step forward in general decision-making processes can be expected.

## Acknowledgement

The authors would like to thank the Society of Actuaries for granting the access of the database for RP-2000 Mortality Tables.

## References

1) Berry, Michael J. A. and Linoff, Gordon S. (2000) *"Mastering Data Mining"*, New York, N.Y.: Wiley.

2) Bishop, C.M. 1995. Neural Networks for Pattern Recognition. New York: Oxford University Press.

3) Breiman, Leo, Friedman, Jorome H., Olshen, Richard A. and Stone, Charles J. (1984) *Classification and Regression Trees.* New York, N.Y.: Chapman & Hall.

4) Cheeseman, P., Kelly, J., Self, M., Stutz, J., Taylor, W., and Freeman, D. 1988. AutoClass: ABayesian classification system. In 5th Int'l Conf. on Machine Learning, June, Morgan Kaufman.

5) GAVRILOV, L.A., AND GAVRILOVA, N.S. 1991. *The Biology of Life Span: A Quantitative Approach.* Chur, Switzerland and New York: Harwood Academic Publishers.

6) Goldberg, D.E. 1989. Genetic Algorithms in Search, Optimization and Machine Learning. Morgan Kaufmann.

7) Hand, D., Mannila, H, and Smyth, P. (2001). *Principles of Data Mining.* Cambridge, Massachusetts: MIT Press.

8) HELIGMAN, L., AND POLLARD, J.H. 1980. ''The Age Pattern of Mortality," *Journal of the Institute of Actuaries* 107:49–75.

9) Hosmer, D. W. and Lemeshow, S. (1989), *Applied Logistic Regression.* New York, N. Y.: John Wiley & Sons.

10) Kass, G. V. (1980) *"An Exploratory Technique for Investigating Large Quantities of Categorical Data,"* Applied Statistics, 29:119-127.

11) KEYFITZ, N. 1968. *Introduction to the Mathematics of Population,* Reading, Mass.: Addison Wesley.

12) MCNOWN, R. 1992. ''Modeling and Forecasting U.S. Mortality. Comment,'' *Journal of the American Statistical Association* 419:671–72.

13) MCNOWN, R., AND ROGERS, A. 1989. ''Forecasting Mortality: A Parameterized Time Series Approach,'' *Demography* 26: 645–60.

14) MCNOWN, R., AND ROGERS, A. 1992. ''Forecasting Cause-Specific Mortality Using Time Series Methods,'' *International Journal of Forecasting* 8:413–32.

15) POLLARD, J., AND STREATHFIELD, K. 1979. *Factors Affecting Mortality and the Length of Life,* Number 197. North Ryde, Australia: Macquarie University, School of Economic and Financial Studies.

16) Ripley, B. D., (1996) *Pattern Recognition and Neural Networks,* Cambridge University Press.

17) ROSE, M. 1991. *Evolutionary Biology of Aging.* New York: Oxford University Press.

18) SAS Institute, (2000) *Enterprise Miner*, Cary, N.C.: SAS Institute.

19) Tuljapurkar, S., (1998). "FORECASTING MORTALITY CHANGE QUESTIONS AND ASSUMPTIONS", *North American Actuarial Journal*, Vol. 2, No. 4, pp127-135.

20) Tuljapurkar, S. and Boe, C. (1998). "Mortality change and Forecasting: How much and How Litter do we know?" *North American Actuarial Journal*, Vol. 2, No. 4, pp13-48.

21) WACHTER, K., AND FINCH, C., ed. 1997. *Biodemography of Aging.* Washington, D.C.: National Academy Press.

22) WILKIN, J.C. 1981. ''Recent Trends in the Mortality of the Aged,'' *Transactions of the Society of Actuaries* XXXIII: 11–62.

23) YASHIN, A., MANTON, K., AND VAUPEL, J. 1985. ''Mortality and Aging in a Heterogenous Population: A Stochastic Process Model with Observed and Unobserved Variables,'' *Theoretical Population Biology* 27:154–75.